Vocal Attack Time of Different Pitch Levels and Vowels in Mandarin

∗Ruifeng Zhang, † R.J. Baken, and ∗Jiangping Kong *∗Beijing, China and † Woodstock, New York, USA and Hong Kong University, Hong Kong SAR, China*

Summary: The purpose of this study was to investigate how vocal attack time (VAT) varies when young adults articulate the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch levels. Sound pressure (SP) and electroglottographic (EGG) signals were recorded simultaneously from fifty-three male and fifty-three female subjects saying sustained /A/, /i/ and /u/ at five equally spaced pitch heights, each being higher than the preceding one. Then analyses of means, variance and correlation were performed to explore the relationships of VAT/pitch levels and VAT/vowels. Findings were: As mean STs (semitone) increase linearly from levels one to five, mean VATs decrease nonlinearly in a big group of subjects but increase nonlinearly in a small group of them. Based on the body-cover model of F0 control, data here lead to the guess that different people incline to use different strategies in increasing pitch height. When males, females and males plus females are considered as a whole, average STs and VATs tend to be positively correlated among the three vertex vowels.

Key Words: Vocal attack time—Pitch levels—Vertex vowels—Semitone.

INTRODUCTION

Vocal attack time

Vocal attack time (VAT) is the time lag between the rise of the simultaneously recorded sound pressure (SP) and electroglottographic (EGG) signals, measured at the onset of phonation[1]. When airflow goes through the glottis during the initiation of a vowel or a voiced consonant, the vocal folds oscillate with very small amplitudes before their first contact is achieved and stabilized. Therefore, the SP signal, which is a record of sound pressure emitted from the mouth, begins its growth of amplitude well before the vocal folds touch each other, but the EGG signal, as a record of vocal-fold contact area, has nearly no amplitude until the vocal fold contact occurs, and only after that does its magnitude show up and grow. This is the occasion when VAT values are positive. However, in other cases when the initiation of EGG signals leads that of SP signals, namely, when vocal fold contact precedes the appearance of SP signals such as in a hard glottal attack, VAT is negative. If the two sorts of signals rise at the same point of time, VAT equals zero. Consequently, VAT can be taken to be the duration from the start of vocal-cord oscillation to the instant for the first vocal-cord contact, and provides a useful index to indicate the pre-phonation laryngeal adjustment.

The effectiveness of VAT measurement was experimentally verified by Orlikoff et al[1]. using five vocally normal subjects. EGG and SP signals of different phonation types were recorded synchronously with high-speed video-endoscopy, from which a digital kymogram (DKG) was generated. The DKG attack duration data obtained by hand were then compared with the VAT measures extracted by using computer programs. The strong and direct relationship between the VAT and DKG-measured data proved VAT to be a valid and convenient measure of vocal attack. In 2012, a figure of merit (FOM), which assesses a critical assumption of vocal startup on which the VAT measure is based and therefore represents integrity of the derived measure, was proposed by Roark et al[2]. for the VAT measurement of sustained /a/. SP and EGG signals from 102 tokens were visually inspected to empirically derive a criterion level of FOM less than 0.75 to indicate when the assumption underlying the measurement had failed and the VAT value obtained should be disregarded. The example of using VAT for nonlinguistic research was the measurement by Roark et al[3] acquiring normative data of VAT in healthy young adults. They collected SP and EGG signals from fifty-five males and fifty-seven females performing multiple tokens of three tasks (sustained /a/, “always”, and “hallways”) at comfortable pitch and loudness. The average VATs were significantly shorter for females than for males and mean VAT was 1.98 milliseconds in the screened sample of normal

From the ∗Department of Chinese Language and Literature, Research Center for Chinese Linguistics and Joint Center for Language and Human Complexity, Peking University, Beijing, China; and the † Department of Speech-Language Pathology, New York Medical College, Valhalla, New York.

Address correspondence and reprint requests to R.J. Baken, Department of Speech-Language Pathology, New York Medical College, Valhalla, New York. E-mail: rbaken@hvc.rr.com

young speakers. The use of VAT in linguistic research was exemplified by the measurements done by Ma et al[4]. examining the association between VAT and tone in Cantonese speakers.

Pitch levels and vowels

It is well-known that pitch increases along with the acceleration of vocal fold oscillation, resulting from step-by-step augmentation of the vocal folds' tension. The VAT study of three phonation types by Orlikoff et al[1]. seems to suggest that tenser vocal folds tend to be associated with smaller VAT values. Therefore, how VAT varies with increasing pitch in Mandarin Chinese appears to be an attractive research subject that has never been touched upon. We required the subjects to produce vowels at five different linguistically unconstrained pitch levels out of two considerations. For the purpose of devising tone-letters, Chao[5] divided the pitch range of a person into four equal parts with five points numbered 1, 2, 3, 4, 5, corresponding to low, half-low, medium, half-high and high respectively. It has also been found by subsequent linguistic researchers that no language uses over five pitch levels to distinguish its tones[6]. On the other hand, many subjects felt it natural and easy to space five pitch levels equally within their voice range, but difficult to manage six or more levels that way. So the present study, by focusing on five sustained pitch heights that are not linguistically distinctive, is intended to be conducive to future work on language tones. The three vertex vowels /A/* /i/ and /u/ in Mandarin Chinese were chosen to be produced at five pitch levels, because they occupy the utmost points on the vowel chart and represent the entire scope of tongue movement during articulation. All in all, the purpose of this study is to explore how VAT varies when young people produce the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch heights.

METHOD

Subjects and instrumentation

Fifty-three females (18 to 22 years old) and fifty-three males (18 to 22 years old), all of whom were college students, participated in the research. They spoke standard Mandarin Chinese for daily communication, had no voice or hearing problems, and were all in good health at the time of recording. The recording was accomplished in the sound-treated booth at the Language Laboratory of the Chinese Department, Beijing University, where the background noise was below 25dBA. The Adobe Audition 2.0 on the computer (Lenovo, x220i) was set at the stereo interface with a sampling rate of 44100 Hz and a resolution of 16 bits for each channel. The electroglottograph (Model 6103) used for collecting EGG signals and the microphone and sound card (Creative Labs Model No.sb1095) used to gain SP signals were synchronously connected to the computer through a sound console (Behringer XENYX502). With their lips about 10 cm away from the microphone, the subjects were asked to say sustained /A/, /i/ and /u/ at five pitch heights, each of them being higher than the preceding one. All pitch levels were repeated twice and 30 tokens (3 × 5 × 2) were obtained from each speaker.

Parameter extraction

F0, VAT and FOM measures were extracted largely automatically from the speech samples using the software developed by Roark et al[3], which processed signals in four stages: signal verification, signal segmentation, F0-based frequency filtering and signal modeling, and extraction of measures. From the 3180 samples (1590 for males and 1590 for females), 3165 values (1590 for males and 1575 for females) were obtained for each of the three parameters, with 15 female speech recordings unable to be evaluated by the software, possibly due to the poor quality of their EGG signals.

Data preprocessing

Since our research required the subjects to pronounce each of the three vowels with increasing pitch heights, the F0 values they produced for each vowel should theoretically increase with the level shift from one to five. Consequently, a correlation analysis was firstly done that discarded 210 ineffective speech samples (140 for females and 70 for males) whose pitch values had a negative correlation with the pitch level numbers. The 2955 measures left were then divided into ten groups: male-level1, male-level2, male-level3, male-level4, male-level5, female-level1, female-level2, female-level3, female-level4 and female-level5. Each group was processed separately in the same way: Since vocal fold vibrations normally do not go beyond 500Hz, measures that were beyond ±3 standard deviations from the mean F0 were firstly removed from each group; And secondly, according to the observation that there

*Since /A/ in Mandarin is the lowest central vowel and different from the front /a/ and back /ɑ/ on the IPA chart, Chinese linguists prefer to represent it with a capital letter.

were more outliers among VAT values, measures that were beyond ±2 standard deviations from the mean VAT were eliminated. Among the 2827 measures that remained eventually, F0 ranged from 77.3Hz to 497.06Hz, the mean (SD) being 220.35 Hz (75.26 Hz), and VAT from -56.26ms to 53.13ms, the mean (SD) being 0.75 ms (8.69 ms). For the Excel and SPSS analyses below, all F0 values were converted to semitones (ST) re 64.66 Hz. This reference level was chosen not only because it was close to the minimum pitch value 77.3 Hz but also because Liu[7] had been using it in his groundbreaking research of Mandarin tones.

RESULTS

VAT and pitch levels

Among the 1458 male speech samples, pitch ranges from 3.09ST to 26.06ST (mean = 15.72 ST; SD = 4.19) and VAT from -25.67 ms to 39.27ms, (mean = 0.91 ms; SD = 7.46). Among the 1369 female speech samples, pitch ranges from 16.45ST to 35.31ST (mean = 24.98ST; SD= 3.49) and VAT from -56.26 ms to 53.13 ms (mean = 0.59ms; SD = 9.84ms). One-way analyses of variance and post hoc tests have shown that, at significance level p = 0.01 for both males and females, ST values are significantly different between any two of the five pitch levels ($p < 0.01$ for all), exceeding a preselected $\alpha = 0.01$, while for VAT, significant differences are only seen between pitch level one and each of these levels: two, three, four and five ($p < 0.01$ for all). According to a correlation analysis, ST has a significant negative correlation with VAT among all the 2827 speech samples (N=2827, r = -0.077, $p < 0.01$). When male and female measures are calculated separately, the significant negative correlation still holds for both but with a higher strength of correlation for females than for males (N=1458, r= -0.084, $p < 0.01$ for males; N=1369, r= -0.115, $p < 0.01$ for females).

In Table 1 are listed the maximum, minimum, mean and standard deviation of ST and VAT across pitch levels calculated according to this grouping: A: males (1458 tokens); B: females (1369 tokens); and C: total (2827 tokens). Here, vowel considerations and individual differences are temporarily left aside. See Table 1. As pitch levels shift from one to five, all three groups show a linear increase of pitch, but a nonlinear and non-monotonic decrease of VAT: unlike ST means, each mean value of VAT from Levels Two to Five is not always larger than the one that follows; However, in all these cases, the average VAT at Level One is always the largest among the five pitch levels, and is much larger than the mean VATs at Levels Two, Three, Four and Five, making all five going on a downward trend. Mean STs and VATs as a function of pitch levels for the three groups are indicated in Figure 1, from which VAT variations, compared with ST changes, can be seen more intuitively: From Level One to Levels Two and Three, there is a deep declination, and then a clear upturn at Level Four, followed by a steep dip at Level Five to the minimum mean VAT value.

TABLE 1. Maximum, minimum, mean and standard deviation of ST and VAT across pitch levels in Groups A: males, B: females and C: total.

groups		A: males (N=1458)				B: females (N=1369)				C: total (N=2827)			
pitch levels		max	min	mean	SD	max	min	mean	SD	max	min	mean	SD
1	pitch(ST)	18.87	3.09	11.46	3.09	26.40	16.45	20.75	1.93	26.40	3.09	15.96	5.32
	VAT(ms)	39.27	-21.11	2.69	8.64	53.13	-46.24	3.33	12.26	53.13	-46.24	3.00	10.54
2	pitch(ST)	21.37	5.19	14.25	3.20	29.37	17.97	23.46	1.94	29.37	5.19	18.74	5.32
	VAT(ms)	22.83	-22.70	0.59	7.35	27.73	-36.74	-0.27	8.82	27.73	-36.74	0.17	8.10
3	pitch(ST)	23.03	7.91	16.15	3.32	31.44	19.55	25.22	2.10	31.44	7.91	20.53	5.33
	VAT(ms)	20.45	-21.72	0.39	6.56	31.93	-29.37	0.07	8.18	31.93	-29.37	0.24	7.38
4	pitch(ST)	23.40	8.73	17.70	3.24	34.07	21.02	26.94	2.49	34.07	8.73	22.21	5.45
	VAT(ms)	20.11	-21.22	0.59	7.00	26.44	-23.61	0.95	7.69	26.44	-23.61	0.77	7.34
5	pitch(ST)	26.06	11.01	19.11	3.26	35.31	23.26	28.66	2.36	35.31	11.01	23.68	5.57
	VAT(ms)	15.85	-25.67	0.27	7.36	27.57	-56.26	-1.20	10.98	27.57	-56.26	-0.43	9.29

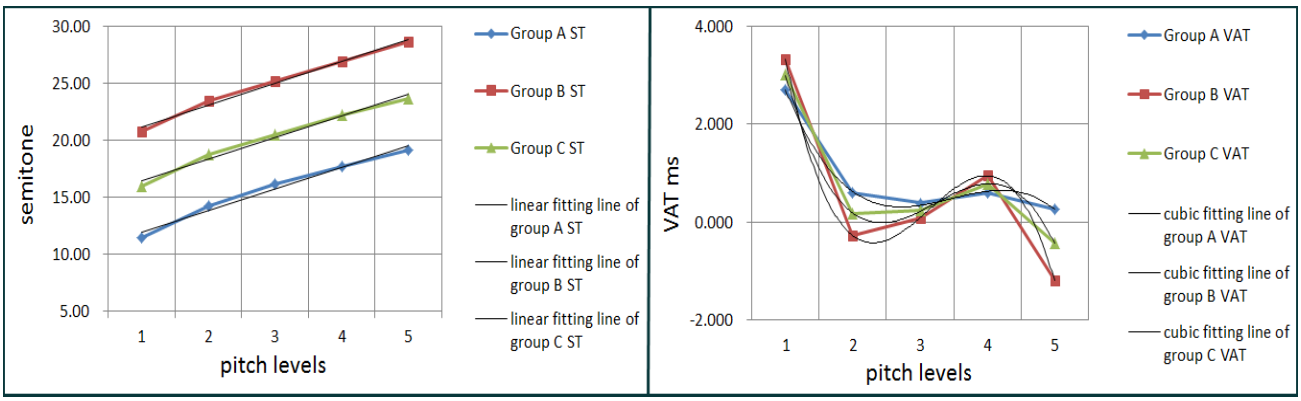

FIGURE 1. Average pitch in semitones and mean VAT in milliseconds as a function of pitch levels for Group A: males (N=1458), B: females (N=1369), C: total (N= 2827).

Now consider how individuals uttered five pitch levels differently. Because each two repetitions of a pitch level were produced consecutively (e.g. /A/level 1→/A/level 1→/A/level 2→/A/ level 2→/A/level 3→/A/ level 3→/A/ level 4→/A/ level 4→/A/ level 5→/A/ level 5, etc.) and the pitch values of them were nearly the same, they can and have to be averaged if we want to see inter-subject differences associated with ST-VAT correlation while ignoring the slight differences between tokens repeated by the same person. The findings after that done are displayed in Figure 2. Among the 103 subjects (52 males and 51 females) whose data were retained for analyses, 36 (35%) produced all /A/, /i/ and /u/ with negative VAT-ST correlation coefficients, 25 (24%) uttered two of the three vowels with negative VAT-ST correlation coefficients and one with positive ones, 20 (19%) pronounced one vowel with VAT-ST negatively correlated and two with them positively correlated; 11 (11%) articulated all three vowels with positive VAT-ST correlation coefficients. The remaining 11 subjects (11%) cannot enter any of these categories because the measures of one or two of their vowels were culled out during data preprocessing. In summary, many subjects produced increasing pitch heights with VAT declining, but a small number of them did it with VAT increasing, and of course there were also subjects belonging to the intermediate type.

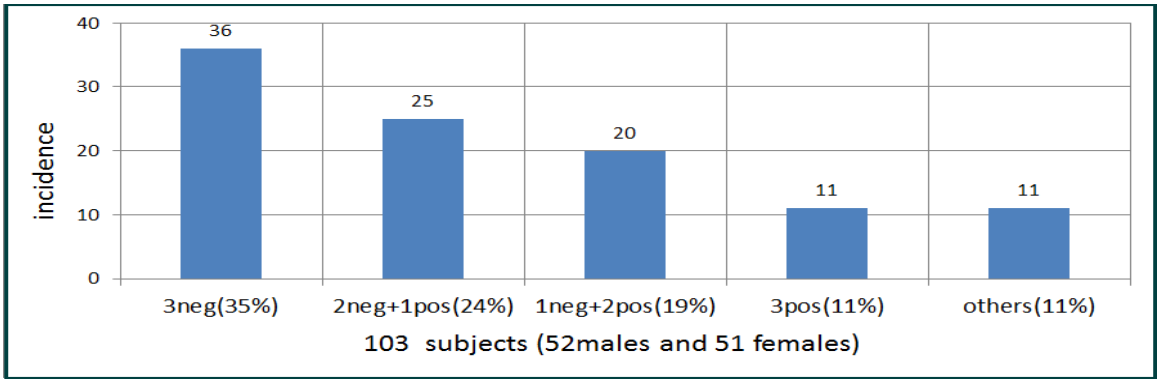

FIGURE 2. The 103 subjects are divided into four categories depending on how many of their three vowels were uttered with negative VAT-ST correlation coefficients and how many with positive ones. 3neg means all three vowels were produced with negative coefficients and 3pos all three with positive ones. 2neg + 1pos and 1neg + 2pos can be understood by analogy.

In Table 2 are listed the means and standard deviations of ST and VAT across five pitch levels of the 36 subjects who uttered all three vowels with negative VAT-ST correlation coefficients and the 11 who did all three with positive ones. Their mean STs and VATs as a function of pitch levels are indicated in Figure 3. The VAT-ST co-variation of the former can be seen in Figure 3 (a): As pitch levels shift linearly from one to five, the mean VATs drop gradually except that there is a slight upturn on pitch level four. But that of the latter shows a quite different picture: As pitch levels go linearly from one to five, average VATs increase gradually except for the small turns at pitch levels two and four (Figure 3b).

TABLE 2. Means and standard deviations of ST and VAT across five pitch levels in the two different categories of subjects.

	36 subjects who uttered all 3 vowels at 5 pitch levels with negative VAT-ST correlation coefficients				11 subjects who uttered all 3 vowels at 5 pitch levels with positive VAT-ST correlation coefficients			
pitch level	mean ST	SD of ST	mean VAT	SD of VAT	mean ST	SD of ST	mean VAT	SD of VAT
1	16.57	5.04	6.29	9.57	16.55	3.63	-1.55	6.03
2	19.05	4.93	0.68	6.63	19.54	3.66	-1.55	5.66
3	20.93	5.02	-0.71	7.33	21.19	3.63	1.29	4.60
4	22.57	5.23	-0.74	7.28	22.43	3.50	3.01	4.05
5	24.25	5.52	-2.93	10.29	24.42	3.39	3.32	4.56

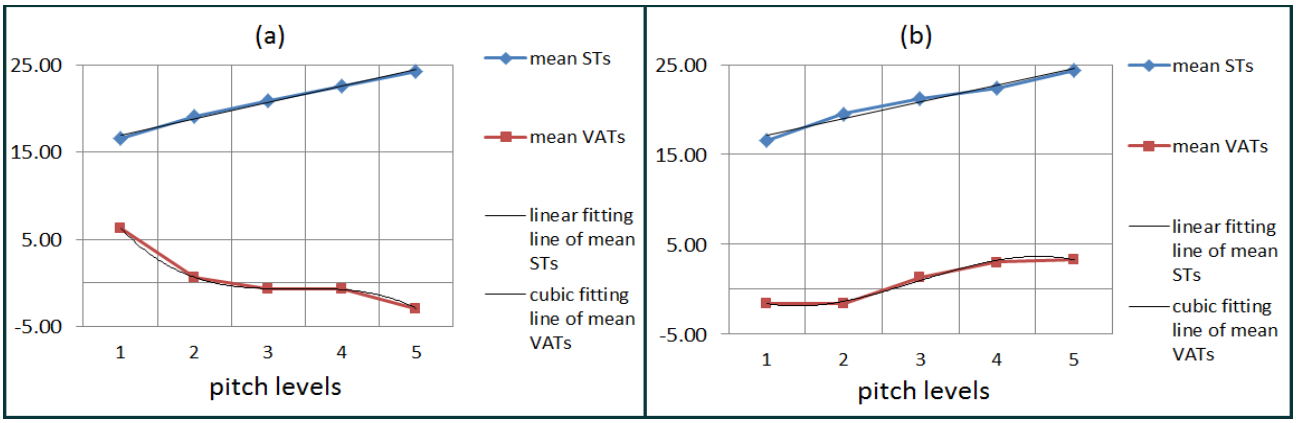

FIGURE 3. Mean STs and VATs across five pitch levels of (a) the 36 subjects who uttered all three vowels with negative VAT-ST correlation coefficients, (b) the 11 subjects who did all three with positive VAT-ST correlation coefficients.

VAT and vowels

To have a picture of VAT variation among different vowels while leaving pitch level considerations aside, mean VATs and STs of /A/ /i/ /u/ are calculated for Groups A: males (1458 tokens), B: females (1369 tokens), and C: total (2827 tokens). The resultant means of the three groups are listed in Table 3 and displayed in Figure 4, from which it is seen that the average STs of /A/ /i/ and /u/ in Groups A, B and C are all thus ordered: /u/> /i/ > /A/, although they are only slightly different between one another. The mean VATs of these three groups are all thus patterned also: /u/>/i/>/A/, with the average VATs of /u/ being the largest among the three vowels and much larger than those of the other two vowels. High vowels tend to have both larger pitch values and longer VATs than low vowels.

TABLE 3. Means of ST and VAT across the three vowels in Groups A: males, B: females and C: total.

groups	A: males		B: females		C: total	
number	N= 1458		N= 1369		N= 2827	
vowels	mean ST	mean VAT	mean ST	mean VAT	mean ST	mean VAT
/A/	15.41	0.07	24.87	-1.39	19.87	-0.62
/i/	15.79	0.87	24.97	-0.15	20.32	0.37
/u/	15.97	1.78	25.08	3.13	20.41	2.44

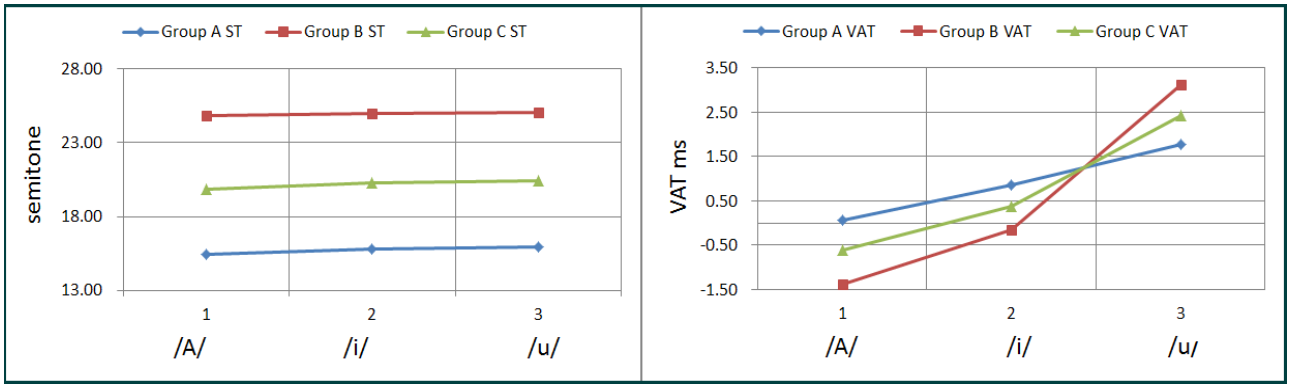

FIGURE 4. Average pitch in semitones and mean VAT in milliseconds as a function of vowels for Group A: males, B: females, C: total.

DISCUSSION

All the analyses concerning VAT and pitch levels point to the same finding. As is indicated in Figure 1, when all males, females or males plus females are considered as a group, a nonlinear contra-variant relationship can be seen between the average values of VAT and ST at five pitch levels, because the best trend line of the increasing pitch is linear while that of the decreasing VAT is cubic. When individuals are taken into consideration, a slightly different picture appears: Many people tend to produce the three vertex vowels (/A/ /i/ /u/) with negative VAT-ST correlation coefficients, but a small number of subjects incline to produce them with positive ones (see Figures 2 and 3). However, the nonlinear relation still holds in both cases, because although Figure 3 (a) presents a contra-variant VAT-ST relationship but Figure 3 (b) shows an orthokinetic one, the best fit lines of the increasing pitches in both graphs are linear while those of the decreasing or increasing VAT are still cubic. The author listened and compared the subjects who pronounced five pitch heights of a vowel with a significantly negative VAT-ST correlation and those who did it with a significantly positive one and his impression was that the former sounded much more laborious than the latter at pitch level five.

Watson et al[8] reported that adjusted mean VAT for their high frequency condition was smaller than the adjusted mean VAT for their mid and low frequency conditions and that VAT appears to be sensitive to increases in vocal fold tension in normal speakers. The present findings not only accord with their statements but also suggests different strategies people tend to use in increasing pitch height. The body-cover model of F0 control proposed by Titze[9] concerns the activities of cricothyroid (CT) and thyroarytenoid (TA) muscles. The contraction of CT elongates the vocal folds but that of TA tends to shorten them. In combination, the two muscles are responsible for most of the length change that can be achieved. TA muscle activity is also used to regulate the effective depth of vocal fold vibration, which reduces rapidly as pitch increases. At low to intermediate F0 and loud productions, namely, in modal voice conditions, CT and TA activities are both relatively low and a significant portion of the vocal fold body is in vibration while the mucosa and ligament remain somewhat lax. A rise in F0 is generally obtained by increased TA activity, as long as CT activity is not near its maximum. As pitch boosts from high to falsetto or a high F0 is to be achieved (as in high-pitched singing), CT activity gradually becomes dominant while the TA action is decreased. Only the surface of the vocal fold vibrates, with a stiff ligament and a loose mucosa in combination[9]. Some guesses can be made based on this model of pitch control and the data reported here. Many people are not good at using falsetto, and, when required to utter vowels at five increasingly higher pitch levels, tend to start at the very low point and go step by step through modal voice register with the vocal fold tension being increased gradually. But quite a few speakers, untrained but skillful in using voices, prefer to start at a relatively higher point and gradually get somewhere around falsetto where the vocal folds, on the contrary, turn somewhat slack. This might be the reasons why the former displayed a negative VAT-ST correlation but the latter a positive one.

It was reported by Zhang[10]and Dong[11] that the three Chinese vertex vowels, while pronounced comfortably, display their intrinsic F0 in such a pattern: /u/>/i/>/A/. The results here support their finding. Figure 4 indicates that both mean STs and mean VATs of /A/ /i/ and /u/ in men, women, and all subjects combined are all ordered: /u/> /i/ > /A/, suggesting that VAT of the three vowels normally tends to be positively correlated with their intrinsic pitch. But what causes such a relationship needs to be further explored.

CONCLUSION

The purpose of this study was to investigate how VAT varies when young adults articulate the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch levels. In a large group of young adults it was found that pitch and VAT change in opposite directions, but in some individuals the two vary in the same direction. But in all cases, pitch and VAT tend to present a nonlinear relationship. The possibility is that people, out of habit or by reason of laryngeal physiology, tend to manipulate their vocal folds in two different ways when they increase pitch from low to high levels. Those who utter vowels at increasing pitch levels with positive VAT-ST correlation coefficients may have more potential in using falsetto, which of course needs to be further investigated. Following this research, our study on VAT of the lexical tones in Mandarin Chinese is now under way.

Acknowledgements

Thanks go to all the voice experts for their gentle participation of the investigation. This research was funded by the National Natural Sciences Foundation of China and the grant number was 61073085.

REFERENCES

1.Orlikoff RF, Deliyski DD, Baken RJ, Watson BC. Validation of a glottographic measure of vocal attack. *J Voice*. 2009; 23:164-168.

2.Roark RM, Watson BC, Baken RJ. A figure of merit for vocal attack time measurement. *J Voice.* 2012; 26:8-11.

3.Roark RM, Waston BC, Baken RJ, Brown DJ, Thmas JM. Measures of vocal attack time for healthy young adults. *J Voice.* 2012; 26:12-17.

4.Ma EP-M, Baken RJ, Roark RM, Li P-M. Effect of tones on vocal attack time in Cantonese speakers. *J Voice.* 2012; 26: 670.e1-670.e6.

5.Chao Y R. A system of “tone-letters”. *方言*.1980; 2: 81-83.

6.Maddieson I. Universals of tone. *Universals of human language*. 1978; Volume 2: 338.

7. 刘复. 乙二声调推算尺. *史语所集刊*. 1934；4 本 4 分：355-361.

8.Watson BC, Baken RJ, Roark RM, Reid S, Ribeiro M, Tsai W. Effect of fundamental frequency at voice onset on vocal attack time. *J Voice.*2013; 27: 273-277.

9. Titze IR. *Principles of voice production*. 2nd ed. USA: National Center for Voice and Speech; 2000: 211-242.

10.张家騄. 元音的内在基频与讲话方式对共振峰的影响.*声学学报*. 1989 ;14: 401-406.

11.董倩倩. 汉语普通话元音音高再探.*文教资料*. 2010: 27-28.